ERIC Document Resume	See Instruction Sheet	DATE OF RESUME 6/24/65	DATE RECEIVED

A. PUBLICATION DATA

1. ERIC NUMBER ED001274

2. ORIGINATOR CONTROL NO(S). NA

3. IDENTIFICATION NO. CE 1993

Sutton, Joseph T. and Allen, Eliot D.

5. TITLE OF DOCUMENT ABSTRACTED: The Effect of Practice and Evaluation on Improvement in Written Composition

6. PUBLICATION TITLE: N.A.

7. EDITOR(S): N.A.

8. PLACE OF ORIGIN

ORG. Stetson University

ST. Deland, Florida

CITY, STATE

TELE. EXT

INST/LOC CODE

9. PUBLISHER OR SOURCE

Cooperative Research, Office of Edu., Dept. of HEW, Washington, D. C.

10. VOL.	11. PAGES	12. MO./YR.	13. TOTAL PP.
N.A.	N.A.	1964	99

DO NOT WRITE IN SPACE BELOW

MICROFILM COPY AVAILABLE YES ☒ NO ☐

COPYRIGHT GRANTED YES ☐ NO ☒

COPYRIGHT RELEASE FILED YES ☐ NO ☒

TYPE OF REPORT
PROPOSAL ☐
IN-PROGRESS ☐
FINAL ☒
FOLLOW-UP ☐

OTHER ERIC REFERENCES:

ERIC SATELLITE CODE

B. INDEXING DATA

14. SUBJECT AREA

15. TYPE OF ENDEAVOR Research

16. INDEXING TERMS (6-12)

	Student evaluation	Writing practice	
Composition	Teacher evaluation		
Composition skills	Teaching techniques		
Evaluation difficulties	Writing ability		

17. ABSTRACT (MAX. 250 WORDS)

The problem of teaching written composition has several facets. Writing and reading themes are time-consuming and costly activities. Increasing enrollment in schools and colleges aggravate these difficulties. Controlled research has not demonstrated that practice in writing or careful grading of themes contributes significantly to the improvement of student writing. Most teachers, nevertheless, feel that students should write more and that their papers deserve and require careful attention.

The experiment discussed in this research report undertook to test three hypotheses. The first hypothesis was that writing practice followed by evaluation is significantly more effective in improving the writing ability of college students than no writing practice (and consequently no evaluation). Second, evaluation by peers is as effective as evaluation by teachers in influencing the improvement of student compositions. The final hypothesis was that direct practice in the evaluation of papers written by peers will result in significantly greater improvement in composition than will practice in writing for evaluation by others.

To test the three hypotheses two types of data were obtained: scores from objective tests and evaluations of written compositions. A detailed explanation of the procedures followed in this experiment, the compiled data, and the results of the study are presented.

In addition to the three major hypotheses, several areas of related research are covered. The related areas include the relationship between frequency of writing and improvement in writing, difficulties and inadequacies in the evaluation of written composition, and the effect upon writing ability of evaluation by or of peers.

ERIC No.

18. SOURCE OF ABSTRACT

C. FUNDING DATA

19. SUPPORTING ORGANIZATION

20. DIVISION OR DEPT.

21. CONTRACT NO.

PROJECT NO.

22. TOTAL COST OF PROJECT

23. DURATION OR PROJECT

U.S. Department of Health, Education, and Welfare
Office of Education

ERIC Form 1

ED001274

C.R.P.

NO. 1993

SUTTON

and

ALLEN

STETSON
UNIVERSITY

THE EFFECT OF PRACTICE AND EVALUATION ON IMPROVEMENT IN WRITTEN COMPOSITION

COOPERATIVE RESEARCH PROJECT NO. 1993

JOSEPH T. SUTTON and ELLIOT D. ALLEN

STETSON UNIVERSITY
DELAND, FLORIDA

REPRINTED BY PERMISSION
U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN PRINTED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

THE EFFECT OF PRACTICE AND EVALUATION
ON
IMPROVEMENT IN WRITTEN COMPOSITION

Cooperative Research Project No. 1993

By

Joseph T. Sutton

and

Eliot D. Allen

STETSON UNIVERSITY

DELAND, FLORIDA

1964

The research reported herein was supported by the Cooperative Research Program of the office of Education, U. S. Department of Health, Education, and Welfare.

TABLE OF CONTENTS

Chapter

Page

ACKNOWLEDGEMENTS ii

I. THE PROBLEM, OBJECTIVES, AND RELATED RESEARCH 1

II. PROCEDURE 11

III. RESULTS 20

IV. DISCUSSION AND CONCLUSIONS 49

BIBLIOGRAPHY 60

APPENDIX A. Theme Topics 63

APPENDIX B. Rating Sheet 73

APPENDIX C. Summary of Data from All Students . . 75

APPENDIX D. Analysis of Variance Tables Not Included in the Text. 81

ACKNOWLEDGEMENTS

The experiment described in this report was made possible by the cooperation and joint support of the U. S. Office of Education and Stetson University. In addition, the encouragement, assistance, guidance, and labor of many individuals contributed essentially to its completion. Advice and direction were given to us on occasion by Dr. Paul Diederich, of the Educational Testing Service; Dr. Benjamin Rosner, of Brooklyn College; Dr. Maurice Tatsuoka, of the University of Illinois; and Dr. Stanford Ericksen, of the University of Michigan. The work of Dr. Warren Jones and Dr. William Taylor, of Stetson University, as research associates, was indispensable to the carrying out of the experiment. Notable for their patience and fidelity during an extended task were our five graders: Dr. Byron Gibson, Dr. Ann Morris, and Mr. Richard Langford, all of Stetson University; and Miss Carolyn Damianos, of Yale University; and Miss Claudine Walker, of the University of Florida. Student assistants were Mrs. Lynn Sawyer, Miss Katrina Hench, Mr. Louis Phillips, and Mr. James Bicklehaupt. Mr. Vernon Cope rendered many valuable services in computer programing and data processing. Finally, without the assistance, voluntary or otherwise, of the one hundred and seventeen Stetson students who were the subjects, the experiment could not have proceeded. To all these, we extend our thanks and appreciation.

Joseph T. Sutton
Eliot D. Allen

LIST OF FIGURES

Figure Page

1. Experimental Design 12

2. Line Graph Showing General Decline in Theme Rating over the Semester 47

LIST OF TABLES

Table Page

1. Means and Variabilities of Two Objective Tests of English Composition Ability 21
2. Analysis of Variance of Scores on the College Entrance Examination Board English Test . . . 22
3. Analysis of Variance of Scores on the Cooperative English Test (English Expression) 23
4. Means and Standard Deviations for Each Section on Total Rating Sheet Scores 25
5. Analysis of Variance of Mean Raw Rating Scores Assigned to the Writing Samples 26
6. Means and Standard Deviations for Each Section for Rankings of the Six Papers Written Before and the Six Papers Written After the Experiment 27
7. Analysis of Variance of the Pooled Ranks of All Graders for Writing Samples 28
8. Means and Standard Deviations of Rating Assigned to All Subjects by Each Grader . . . 29
9. Means and Standard Deviations of Standardized Total Ratings 30

Table Page

10. Analysis of Variance of the Standardized Ratings 31

11. Means and _F_-Ratios of the Best Papers Written Before and After the Experiment 34

12. Means and Standard Deviations of Raw Ratings for Each Grader for Each Factor . . . 36

13. Means and _F_-Ratios of Pooled Standardized Summary Scores on the Five Sub-Sections of the Rating Sheet 38

14. Intercorrelations among the Rating Sheet Factors and the Objective Tests 40

15. Intercorrelations of the Total Ratings and Sub-Total Rating Factors of 72 Papers with Ratings Assigned upon Regrading by Each Grader 44

16. Mean Rank-Order Correlation of Six Sets of Twelve Papers Ranked on Two Occasions by Each Grader 45

17. Mean Correlations between Ratings and Rankings for the Same Papers as Judged by Each Grader 46

18. Mean Ratings per Order Assigned by All Graders to All Papers Written by All Students 47

Table Page

19. Mean Rating by All Graders on All Students for Each Topic 48

20. Analysis of Variance for the Pooled Standardized Ratings on the Best Paper Written before and the Best Paper Written after the Experiment 82

21. Analysis of Variance for the Mean Pooled Standardized Ratings on the Best Two Papers Written before and the Best Two Papers Written after the Experiment 83

22. Analysis of Variance of the Mean Pooled Standardized Ratings of the Best Three Papers Written before and the Best Three Papers Written after the Experiment 84

23. Analysis of Variance of the Mean Pooled Standardized Ratings of the Best Four Papers Written before and the Best Four Papers Written after the Experiment 85

24. Analysis of Variance of the Mean Pooled Standardized Ratings of the Best Five Papers Written before and the Best Five Papers Written after the Experiment 86

25. Analysis of Variance of the Mean Pooled Standardized Ratings of the Ideas Factor 87

Table Page

26. Analysis of Variance of the Mean Pooled Standardized Ratings of the Mechanics Factor 88

27. Analysis of Variance of the Mean Pooled Standardized Ratings of the Wording Factor 89

28. Analysis of Variance of the Mean Pooled Standardized Ratings of the Form Factor 90

29. Analysis of Variance of the Mean Pooled Standardized Ratings of the Flavor Factor 91

CHAPTER I

THE PROBLEM, OBJECTIVES, AND RELATED RESEARCH

-i-The Problem

The central problem of this exercise was to study the effect of frequent writing practice on writing improvement as measured by sample compositions. The subsidiary problem was to explore the effect of a variety of evaluation techniques upon improvement in composition, including practice in writing for evaluation by the teacher; practice in writing for evaluation by peers; and no writing, but practice in the evaluation of peer compositions.

The problem of teaching written composition has several facets. Writing and reading themes are time-consuming and costly activities. Increasing enrollments in schools and colleges aggravate these difficulties. Teacher evaluation of written themes is notoriously unreliable (8 and 24). Furthermore, controlled research has not demonstrated that practice in writing or careful grading of themes contributes significantly to the improvement of student writing (14). Most teachers and curriculum authorities, nevertheless, feel that students should write more and that their papers deserve and require careful attention. If writing practice *per se* is found to be effective, then English teachers need not be distracted by the perennial reminders that their evaluations of compositions are not psychometrically sound. They may then have the assurance that requiring pupils to write is good teaching even if the evaluation of ordinary daily or weekly compositions lacks the refinements required for careful research. If, however, writing ability is not significantly advanced by writing practice, research efforts may be intensified to find other types of exercises that will prove to be more effective.

An alternative or supplement to teacher evaluation of compositions is evaluation by peers. A frequent suggestion made by teachers is that pupils exchange papers for criticism or grading, or that they sit in small groups and review each other's work (5, 9, 13, 18, 23). This process may help to reduce the time which the teacher needs to give to each paper, but no one now assumes that it is a substitute for teacher evaluation. Furthermore, there do not appear to be any controlled studies demonstrating the effectiveness of peer evaluation.

-ii-Objectives

The present experiment undertook to test three hypotheses. The first hypothesis was that writing practice followed by evaluation is significantly more effective in improving the writing ability of college students than no writing practice (and consequently no evaluation). Second, evaluation by peers is as effective as evaluation by teachers in influencing the improvement of student compositions. The final hypothesis was that direct practice in the evaluation of papers written by peers will result in significantly greater improvement in composition than will practice in writing for evaluation by others.

These hypotheses were necessarily somewhat conjectural, since there are no well-articulated theories of written composition from which hypotheses may be rigorously deduced. Two major considerations influenced their selection. The first was the remarkable absence, regarding these fundamental issues, of empirical evidence that might be used to guide educational practice. Second, there is within the systematic psychology of learning a long tradition concerned with the functional relationship between behavior and its consequences (*e.g.*, in this case, evaluation). Within the framework of stimulus-response psychology, evaluation of behavior may be conceptualized in terms of its cue, reinforcing, and energizing effects. As these concepts apply to the evaluation of composition, the teacher's marks and comments on a paper may, in the first instance, perform a cue or informational function by directing the student's attention to desirable responses. In the second case--reinforcement--the teacher's marks may determine the responses that will be strengthened. In the third case--energizing--the value judgment implied in grading the compositions may have the motivational function of energizing or increasing the general level of activity and effort on the part of the student to improve. Thus there was theoretical as well as popular support for the hypothesis that writing practice followed by evaluation would be effective in improving writing ability, though there was little careful experimental evidence for this hypothesis.

Regarding the distinction between the value of peer evaluation and teacher evaluation in the second hypothesis, it seemed to be a reasonable assumption that in general the evaluation by any individual college student was not likely to be as reliable or as valid as the evaluation by an individual college teacher. However, the demonstrably large amount of error variance in teachers' markings suggests that cumulative peer evaluations might well serve to identify areas of weakness as well as or better than a single teacher and thus effectively perform the cue or informational function of evaluation.

It was also felt, in regard to the second hypothesis, that the reinforcing and energizing effects of peer evaluation might be more potent in their control over the student's desire to improve than the comments of the teacher. While relatively few students will identify with the academic culture of the instructor, the impact of the peer culture on secondary school and college students is well documented.

The theoretical assumption behind the third hypothesis was that practice in critical reading of a variety of papers would, through the processes of comparison and contrast of strengths and weaknesses, contribute directly to the development of editing habits that would generalize to the reader's own writing. The significance of self-editing in composition is implicit throughout Skinner's theoretical text, Verbal Behavior (20).

Further, statements of successful writers about their own writing processes seem to suggest that the creative act is accompanied by a self-editing or evaluating act which, in the very moment of creation, modifies what is being created, without dependence upon the outside evaluation of others. This intimate relationship between evaluation and creation is still imperfectly understood.

The third hypothesis also sprang from the belief that whereas the usual procedures (revision of papers on the basis of marking and comments of others) related errors only to their specific context, practice in evaluating a number of papers written by peers would confront the student with similar errors in numerous contexts. This process should increase the learning of generalizations that would transfer to new contexts.

-iii-Related Research

a. Relationship Between Frequency of Writing and Improvement in Writing

The two most pertinent reports are by Heys (14) and Buxton (3). Heys found that high school students who wrote regularly one theme per week throughout the year, revising each paper on the basis of careful evaluation by the instructor, failed to improve, as measured by the STEP Essay Test, as much as control classes who wrote only three papers during the year but used their time for extra in-class and out-of-class reading. There were no significant differences between the reading group and the writing group, as determined by the evaluation of a single writing sample collected before and after the experiment. The writing samples were scored on a nine-point scale by three trained graders for content, mechanics, and diction. Improvement within both groups was shown, but absolute improvement was very small; and differences between the two groups were even smaller, so that respectable statistical significance was not established.

The dissertation by Buxton, using college freshmen, found that students who wrote sixteen weekly essays during the year failed to improve on the Mechanics of Expression portion of the Cooperative English Test more than students who did not write during the year. The experimental writing group did show significant improvement on the Effectiveness of Expression portion, however. The experimental group was further divided into two sections, one of which received only a short critical comment on each paper, with no class discussion. The other received their papers marked in detail, with a commentary, followed by class discussion and revision of the papers. A single essay test was given at the beginning and end of the experiment for criterion purposes. These tests were scored by two teachers working independently, using a scoring sheet divided into fifteen categories. The writing-with-revision group showed significant improvement over the writing-without-revision group in categories labeled _title_, _introduction_, _sentence variety_, _fluency_, and _diction_. There was no significant gain in the categories of _critical thinking_, _originality_, or _organization of material_.

In a study by Maize (17), a control group of college freshmen wrote fourteen themes while an experimental group wrote approximately twenty-eight. The experimental group showed statistically significant superiority over the control group, but the inadequacy of the pre-test and post-test theme examinations, besides the difference in instructional methods and course content, suggests that the apparent superiority of the experimental group may have been due wholly or in part to causes other than frequency of writing.

b. Difficulties and Inadequacies in the Evaluation of Written Composition

A review of studies relative to the problems of evaluation of written composition is found in Huddleston (15). Older studies by Cast (4), Wiseman (25), and Stalnaker (22) reflect the well-known facts of differences between graders and the desirability of some kind of training for them. Two useful recent studies have been published by Anderson (1) and Diederich, _et al_. (10).

Anderson found significant variance among eight writing samples per student distributed over a four-week period. The variance attributed to three markers was also significant at less than the 1 percent level. However, variance within markers over four marking occasions was not statistically significant. The only interaction effect found to be significant was that of testing occasions versus markers. He concluded from this that it would be possible to select a group of very homogenous and stable markers. This corroborates the observation of Diederich that graders might be selected whose inter-reader reliability would be respectably high.

The Diederich study is unique in its scope. The Educational Testing Service employed 53 distinguished readers, representing six different fields, to grade 300 papers written by college freshmen. Factor analysis of correlations revealed five reader-factors: *ideas*, *form*, *flavor*, *mechanics*, and *wording*, and one test-factor. The test factor was identified with three objective tests: the English Composition Test and the SAT Verbal and SAT Quantitative portions of the College Board Examination. The test-factor was found to be significantly correlated with the reader-factors *mechanics* and *wording*, but uncorrelated with the other three. These results appear to give an affirmative answer to the question put by Anderson as to whether there are any unique attributes of essay or composition ability which are distinct from the less complex language or verbal abilities capable of being reliably measured by objective tests. They also suggest that a careful study of improvement in written composition can and should include evaluations of the writing samples themselves.

As an outgrowth of the study, Diederich's group developed a technique for screening test scorers for agreement with the consensus scores of their 53 graders. This technique consists of ranking 18 papers selected at equal intervals over the range of 300 papers. Readers may be selected who will agree with one another if their rank-order correlation with the ranking of these papers is .60 or better. The authors conclude that, in a research study, it would probably take six papers on different topics, each written in a separate testing session, and each graded independently by four tested readers, before the over-all reliability of the grade on writing ability would reach .90.

A more recent development of Diederich's is a scoring sheet which attempts to reflect the factor weights for use by graders in future essay grading.

c. Effect upon Writing Ability of Evaluation by or of Peers

There appear to be no controlled experimental studies on the variable of peer evaluation. There is no reference to possible distinctions in the effect of being evaluated as opposed to doing the evaluation. Numerous sources suggest the value of students' interchanging their papers with peers for criticism (6, 7, 9, 13, 18, 21, 23). These reports describe the procedure (*e.g.*, the Rutgers or Oregon plans) or advocate on logical grounds the use of such evaluation, or report results only in impressionistic fashion. In the Smith study (21), peer evaluation seems to have been simply one technique tried with a large class, and not independently evaluated.

CHAPTER II

PROCEDURE

-i-The Sample

A sample of 117 subjects was drawn from the 1962 freshman class at Stetson University. These students were randomly assigned to five class sections at registration. The 1962 freshman class (N = 394) had a SAT Verbal mean score of 519.7 with a standard deviation of 94.3. The sample SAT Verbal mean score was 551.2 with a standard deviation of 96.0. An attrition of six subjects occurred during the experiment, leaving a total number who finished the experiment of 112, composed of 47 male students and 65 female students.

-ii-The Design

The experiment was cast in a modified split-plot design (11, pp. 227-232). The experimental paradigm is given in Figure 1. All data were gathered in an initial two-week period and a final two-week period. During these periods all students did exactly the same things. The differences among the groups are to be found in the nature of the activities in the period intervening between the pre-experimental and post-experimental data collection periods.

The pre-and post-experimental periods were similar in that during these intervals all subjects wrote six one-hour

FIGURE ONE

EXPERIMENTAL DESIGN

	SECTION	NO. OF STUDENTS	FIRST TWO WEEKS "BEFORE"	MIDDLE TEN WEEKS DEVELOPMENTAL PERIOD	LAST TWO WEEKS "AFTER"
CONTRAL	I	22	write six criterion themes	(next two weeks) write six criterion themes	end of experiment
	II	21	write six criterion themes	ten weeks of literature no writing	write six criterion themes
EXPERIMENTAL	III	25	write six criterion themes	evaluate four peer papers each week no writing	write six criterion themes
	IV	20	write six criterion themes	write one class theme each week evaluation by four peers	write six criterion themes
	V	22	write six criterion themes	write one class theme each week evaluation by professor	write six criterion themes

themes in class and took objective tests. Thus to test the hypotheses, two types of data were obtained: scores from objective tests, and evaluations of written compositions.

Two objective tests were used: the English Composition Test of the CEEB examination and the English Expression portion of the Cooperative English Test, both from the Educational Testing Service. All the subjects had taken the CEEB examination prior to entering college, and their scores were available. All the subjects had also taken Form 1-C of the Cooperative English Test at the start of the semester previous to the experiment. Following the writing of the "After" themes, all the subjects then took Form HB01 of the CEEB English Composition Test and Form 1-B of the English Expression portion of the Cooperative English Test. Although Section I had finished theme-writing early in the semester, the subjects in this section took the final objective tests at the end of the semester at the same time as the subjects in the other sections.

The topics for the written compositions are shown in Appendix B. Each student wrote on a different topic on each of six different occasions in the two-week initial period and again on six occasions in the two-week final period. None of the papers were returned to the students. All students in each section wrote on twelve identical topics, but the order in which the topics were assigned was counterbalanced within each group.

Sections I and II were designated control sections. The students in these sections wrote no themes for their English classes other than the twelve criterion themes just described. In fact, Section I might be called a control-control section. These students wrote their final set of "After" themes immediately upon completing the block of six "Before" themes. Thus the experiment concluded for them after the first five weeks of the semester. Any writing improvement that may be attributed to the adventitious writing practice or other learning experiences of the college student over the course of a semester may therefore be evaluated by comparison of the scores of Section I with those of Section II.

Section II wrote no themes in the ten-week period intervening between the "Before" and "After" criterion papers; their class time was devoted entirely to the study of literature. Comparison of the data from this group with that from the experimental sections should reveal any improvement attributable to the independent variables of the experiment.

Section III also wrote no themes during the middle ten-week period. On the first meeting of each week, each subject in this section read and evaluated four themes written by his peers in Section IV. During the second period of the week, each student compared his evaluation of one of these papers with the evaluations of the same paper made by three other students during the previous period. At the third class period, each student evaluated the rewritten version of a paper previously evaluated by himself. Each of the ten weeks was a repetition of the process carried out in the first week, save for the fact that papers assigned to each reader were selected from a different set of four writers each week.

In Section IV, the subjects followed a ten-week cycle of activities between the pre- and post-testing periods that began with writing a one-hour paper on the first period of each week. At the second meeting each subject reviewed four copies of his paper evaluated by his peers, and at the third meeting revised and rewrote his paper.

Section V was identical to Section IV except that all of the papers written by this group in the middle ten-week period were evaluated by one of two professors of English.

-iii-General Methodology

In order to make possible the writing, evaluating, and rewriting of the papers from Section IV in such rapid fashion, a staff of student assistants made four duplicate copies of each theme on a Thermofax machine. Names of the writers were blocked out so that neither writers nor graders knew the identity of the others. All graders--both student and faculty--used the same rating sheet for evaluating these themes that was later used by the panel of graders who evaluated the criterion themes.

There was no class discussion of the art composition during the course of the experiment. The sole formal instruction in writing given to the subjects was the rating sheets and editorial comments on the papers of Sections IV and V, and the comparisons of grading procedures among the students of Section III.

All students were given a one-hour orientation period prior to the beginning of the study. They were told that they were participating in an experiment but were given no details about the hypotheses or the rationale of the design. No section was given precise information about its role in the experiment. Every student received a copy of the rating sheet and was instructed in the meaning of the categories; further, every student was given copies of four mimeographed themes which he evaluated, using the rating sheet.

All writing, evaluation, and revision were done in the classroom under supervision of either an instructor or a graduate assistant. No experimental materials were taken out of the room by the student. Assignments were passed out to the students in individual folders at the beginning of the hour, and the folders and compositions were collected at the end. The instructor neither discussed the theme topics nor instructed in writing or evaluation during the class hour.

-iv-Theme Topics

Twelve theme topics were selected for use in the criterion themes, and ten others for use in the themes written by Sections IV and V during the intervening ten-week period. Criteria for their selection included these considerations: each topic must be capable of being discussed at a reasonably specific level within fifty minutes; no special knowledge beyond that all freshmen might be expected to have was required; topics were expected to be of at least moderate interest to college freshmen; instructions for each topic could be presented to the students in simple fashion in approximately half a page of mimeographed material. Copies of the topics as they were presented to the students are found in Appendix A.

The ten topics used during the middle portion of the experiment were given to all the students in the same order. The twelve criterion topics, however, were assigned in random order, but distributed so that while each student eventually wrote on all twelve topics, every topic was being used by some students on any given writing occasion. Thus every topic occurred as frequently as a "Before" topic as it did as an "After" topic among the papers written by the students within each of the five sections.

-v-Evaluation of Compositions

In preparation for the evaluation process, all the criterion themes were typed on mimeograph stencils exactly as written, with all errors preserved. The stencils were carefully compared with the manuscript originals before being duplicated; every effort was made to bring the themes to the graders exactly as they were written. The mimeographed versions were identifiable only by code numbers, so that the panel of graders would not be able to identify the writer, the section, or the order in which the themes were written.

Graders for the experiment were chosen from among experienced college teachers of English and superior students in the graduate program. In accordance with the method described by Diederich (10, p. 45) graders were tested on eighteen papers, standarized by evaluation by his panel of fifty-three distinguished graders. Three instructors and two students were selected. They were given additional themes to be graded with the use of the rating sheet, and brought together for discussion periods to adjust differences of interpretation and evaluation. Each grader correlated at least .60 with the consensus scores of the other four graders prior to the evaluation of the criterion papers.

A copy of the rating sheet used in the experiment appears as Appendix B. Its use was possible through the kindness of Dr. Paul Diederich of Educational Testing Service. The five main categories--ideas, mechanics, wording, form, and flavor--with their thirteen sub-categories were an outgrowth of his factor analysis of the data from his large experiment described in *Factors in Judgments of Writing Ability* (10). Each of the sub-categories carries a rating scale from 1 (lowest 5%) to 5 (highest 5%). In addition, there is a space for comments by the grader at the bottom of each sheet.

Grading of the criterion themes was carried out, during the summer of 1963, in two separate steps.

First, each grader was given twelve sets of 112 themes--each set consisting of all 112 themes on one topic, one by each student. To each theme a rating sheet was attached and filled out by the grader. The twelve sets were arranged in different orders for the five graders, so that no student's themes always appeared early, in the middle, or late in the sets. In addition, a block of six themes found early in each set was repeated near the end, as a means of testing the reliability of the grader.

When the first step of the grading process was complete, each grader was then given 112 separate sets of 12 themes--each set consisting of all the themes by a single student. The grader's task was to place them in order from poorest to best. This step was carried out with new duplicate copies so that the graders had no record of the way they had rated the themes during the first step. No indication was given to the grader, of course, as to the order in which they had actually been written; and the themes were placed in the grader's folders in randomly varied orders. Graders completed one set before beginning another. As in the first grading, the ratings, a measure of the graders' reliability in ranking was assured by the placing near the end of the 112 six sets which were the duplicates of six near the beginning.

CHAPTER III

RESULTS

The results of the experiment are presented in two major parts. First, those data that are pertinent to the testing of the empirical hypotheses are subjected to analysis of variance. Second, several comparisons and relationships are presented that are not directly related to the hypotheses testing but which, nevertheless, may be of interest to those who teach or do research in the area of composition.

-i-Objective Test Scores

The College Entrance Examination Board English Test (CEEB) was given to all subjects before and after the body of the experiment. These data are presented in Appendix C. The means and variability of each of the five sections are presented in Table 1. It will not be noted that in this table the N for each section has been reduced to 20, which represents the number of students in the smallest section. The reduction in number was accomplished by first eliminating those students who did not have a complete data record; the remainder were randomly chosen. This process eliminated the complications of possible disproportionality upon the analysis of variance design. The results of the analysis of variance

TABLE 1

Means and Variabilities of Two Objective Tests of English Composition Ability. Each Given Before and After the Experiment.

N = 20 in Each Section.

Section	College Entrance Exam. Board English Test				Cooperative Eng. Test: English Expression			
	Before		After		Before		After	
	M.	S.D.	M.	S.D.	M.	S.D.	M.	S.D.
I	503.7	93.9	541.5	116.6	166.4	7.9	169.3	8.2
II	487.9	72.9	544.0	81.5	164.4	7.1	167.2	8.2
III	509.5	81.2	556.5	101.3	172.1	7.3	173.4	5.8
IV	519.9	64.3	568.9	83.5	163.3	14.9	168.4	6.5
V	513.5	91.9	545.2	87.8	164.3	9.2	166.9	5.2

are shown in Table 2. The data in Table 1 reflect a gain from the first to the second testing in each of the five sections. This consistent difference is substantiated by the F for Before and After main effect of 32.76 in Table 2. On the other hand, the differences among the means of the various sections were relatively small and cannot be attributed to other than chance variations (F = .34). The F-ratio that is pertinent to testing the hypothesis that the methods of handling the experimental groups resulted in gains in composition ability as measured by the CEEB Test is labeled "Methods."

TABLE 2

Analysis of Variance of Scores on the College Entrance Examination Board English Test Given Before and After the Experiment

Source of Variation	Sum of Squares	d.f.	Mean Square	*F*
Between Ss				
Sections	18,507.0	4	4,627.8	.34
Error Between	1,289,453.0	95	13,573.2	
Within Ss				
Before-and-After	98,301.0	1	98,301.0	32.76**
Method (Sec X B&A)	3,709.0	4	927.2	.31
Error Within	285,031.0	95	3,000.3	
Total	1,695,001.0	199		

**Significant at P<.01 for 1 and 95 d.f.

The Method effect, an interaction between Sections and Before-and-After, is tested to ascertain whether it is tenable to assume that treatment of the various groups affected changes in test scores, from the first to the second testing, in significantly different ways among the groups. The obtained *F*-ratio for methods of .31 is considerably less than the 2.48 required for significance at the five per cent level of confidence. Therefore an hypothesis of no difference between the groups cannot be rejected, and the three experimental treatments cannot be said to have had any effect measured upon the subjects in these groups that would make them different from the control subjects. The fact that all groups showed marked improvement on this test must be interpreted as due to some influence such as systematic differences in difficulty of the two forms of the test employed; incidental learning of all the students between the two administrations of the test; increased sophistication in test-taking ability; or some similar factor.

The English Expression portion of the Cooperative English Test (COOP) was also given before and after the experiment as described in Chapter II. These data for each student are reported in Appendix C. The means and variabilities for each section are reported in Table 1, and the analysis of variance is given in Table 3. It will be

TABLE 3

Analysis of Variance of Scores on the Cooperative English Test (English Expression) Given Before and After the Experiment

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	1,294.7	4	323.7	2.74*
Error Between	11,238.3	95	118.3	
Within Ss				
Before-and-After	278.4	1	278.4	9.27**
Methods (Sec X B&A)	162.5	4	40.6	1.35
Error Within	2,852.1	95	30.0	
Total		199		

*Significant at $P < .05$ for 4 and 95 d.f.
**Significant at $P < .01$ for 1 and 95 d.f.

observed that these results are in large measure similar to those for the CEEB test. A major significant difference is again found between the first and second administrations of the test. The _F_-ratio of 9.27 is not as large as the ratio for Before-and-After with the CEEB, but nevertheless significant at beyond the P = .01 level. Variance attributable to Sections is significant at the 5 per cent level. This difference is difficult to rationalize in view of the random assignment of subjects but is not considered important in changing the overall interpretation of the experiment against the mass of other data. Also, the methods effect is independent of these differences. The interaction of section means with Before-and-After means is not significant. Therefore, we must assume that, as measured by two different and widely used objective tests of composition ability, this experiment failed to demonstrate that the experimental and control treatments had significantly different effects upon the subjects.

-ii-Data from the Writing Samples

The data obtained by having each of five judges evaluate each of twelve writing samples from each of 112 subjects were treated in a number of ways. It will be remembered that the graders read each paper on two separate occasions: first to rate the paper using the special rating form; and later to rank the paper among all twelve papers written by a given student. The data obtained from the first procedure will be referred to as "ratings," and the data from the second evaluation will be known as "rankings." These data were prepared for analysis by reducing the sixty values (five graders times twelve papers) obtained by each method to two summary values for each method. These summary values represent the mean scores on writing performance before and after the imposition of the independent variables. They were obtained for the ratings by adding the total rating scores from each of the six papers written before the experiment began by each student as rated by each of the five graders. This sum was then reduced to a mean value by division by thirty (six papers by five graders). The set of six papers written after the experiment was treated in the same fashion. The Before-and-After means for each student are presented in Appendix C. The means and standard deviations of the Before-and-After scores for each section are given in Table 4, and the analysis of variance results in Table 5.

TABLE 4

Means and Standard Deviations for Each Section on Total Rating Sheet Scores on Six Papers Before and Six Papers After the Experiment

Section	Before		After	
	M	S.D.	M	S.D.
1	38.43	5.09	37.94	5.11
2	39.14	4.20	38.69	4.25
3	39.20	5.06	39.12	5.10
4	36.91	5.13	35.90	5.75
5	37.95	4.87	36.69	2.89

TABLE 5

Analysis of Variance Mean Raw Rating Scores
Assigned to the Writing Samples.
N = 20 for Each Section.

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	209.4	4	52.3	1.20
Error Between	4,159.0	95	43.8	
Within Ss				
Before-and-After	21.4	1	21.4	8.92**
Methods (Sec X B&A)	9.0	4	2.2	.93
Error Within	228.1	95	2.4	
Total	4,627.4	199		

**Significant at $P < .01$ for 4 and 95 d.f.

These tables reveal that the only significant _F_-ratio is the one assigned to Before-and-After. An _F_ this large could have been obtained by chance less than one time in a hundred. The important observation in regard to these data is that the After means were smaller than the Before means. Thus we find a _decline_ in writing performance as measured by this method in every section. The absence of significant F-ratios attributable to Section differences or to Methods makes these data from the writing samples appear similar to the data from the objective tests except for the quite unexpected reversal in the direction of the differences between the Before-and-After means.

There remains the analysis of the ranking data. These data were summarized prior to being subjected to analysis of variance in a manner similar to the rating scores. Mean rankings are presented in Table 6. Analysis of variance of these data are reported in Table 7. These results are similar to those from the ratings except that the F for Before-and-After means is not significantly different from chance at the five per cent level of confidence. Thus these results do not corroborate the analysis from the rating data in presenting significant overall decline in writing proficiency over the course of the experiment. However, again, no significance can be assigned either to differences among the

TABLE 6

Means and Standard Deviations for Each Section for Rankings of the Six Papers Written Before and the Six Papers Written After the Experiment

Section	Before		After	
	M	S.D.	M	S.D.
1	6.54	.80	6.45	.74
2	6.52	.82	6.47	.79
3	6.36	.72	6.63	.78
4	6.73	.82	6.26	.84
5	6.88	.80	6.11	.74

TABLE 7

Analysis of Variance of the Pooled Ranks of all Graders for Writing Samples

Source of Variation	Sum of Squares	d.f.	Mean Square	*F*
Between Ss				
Sections	0.0	4	0.0	0.00
Error Between	0.0	95	0.0	
Within Ss				
Before-and-After	2.5	1	2.5	2.13
Methods (Sec X B&A)	6.7	4	1.7	1.39
Error Within	113.3	95	1.2	
Total	122.5	199		

sections in writing performance or to differences in the ways in which they were treated.[1]

-iii-*Analysis with Standardized Ratings*

The procedure described above for reducing the rating values on each of the twelve papers as graded by each of five readers to two summary scores--*i*.*e*., mean ratings Before-and-After--may be open to some question. In comparing the results from the individual graders, it appears that the means and variances among the graders were noticeably and consistently different, a fact attributable, perhaps, to idiosyncracies of grading behavior. See Table 8 for these data.

---

[1] The *F*-ratio of exactly 0 for Sections in this analysis is a statistical artifact attributable to the nature of the ranking data. It will be remembered that the twelve papers from each student were assigned ranks of one through twelve in the ranking process. Therefore, the total rank score on all papers for each student was identical with the total for every other student. There was, consequently, no variation among students nor among section means in these scores.

TABLE 8

Means and Standard Deviations of Ratings Assigned all Subjects by each Grader--N = 112.

Grader	Mean	S.D.
1	37.52	8.28
2	35.32	4.98
3	40.56	12.44
4	39.33	6.94
5	37.75	6.05
Total	38.10	7.74

Simple means of grades from these five judges would not give equal weight to each grader. The graders with the larger intra-grader variance and/or the lower inter-grader correlations would contribute proportionately more to the variance of the means. Therefore each mean was recomputed after transforming each score on each paper into a standard score equivalent. This was done by converting each rating to a z-score, (_i_._e_., subtracting the grader's mean rating on all papers and dividing the remainder by the standard deviation of all ratings given by that grader). These z-scores were then converted to standard scores having a mean and standard deviation equal to that of the distribution of raw rating scores for all graders (Table 8). This last step makes it easier to compare the effect of the standard score conversion analysis with the raw data analysis than if the score had been left in z-score form. Of course the analysis of variance of these data results is identical with that which would be obtained with the z-score data. Finally, the standardized ratings for each student were summed over all graders for the six papers written before and the six written after and mean values obtained in the manner described for the raw rating data. The summary Before-and-After standard score means for all students in the sample are presented in Appendix C, the section means and standard deviations for these data in Table 9, and the analysis of variance in Table 10.

TABLE 9

Means and Standard Deviations of

Standardized Total Ratings

Group	Before		After	
	M	S.D.	M	S.D.
1	38.54	4.58	38.10	4.48
2	39.23	3.73	38.83	4.08
3	39.28	4.95	39.19	4.82
4	37.19	4.84	36.07	5.59
5	38.14	4.29	36.74	3.80

From a comparison of Table 9 with the raw score means of Table 4 and the analysis of variance results of Table 10 with

TABLE 10

Analysis of Variance of the Standardized Ratings.

Before-and-After Pooled over all Graders.

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	191.8	4	47.9	1.18
Error Between	3,848.6	95	40.5	
Within Ss				
Before-and-After	23.9	1	23.9	10.38**
Methods (Sec X B&A)	11.8	4	3.0	1.28
Error Within	218.6	95	2.3	
Total	4,294.7	199		

**Significant at $P < .01$ for 1 and 95 d.f.

Table 5, it is apparent that no important differences in interpretation of the results of the experiment are needed after this refinement in method. The only significant F-ratio remains the Before-and-After effect, ($F = 10.38$) reflecting the consistent drop in writing performance from the beginning to the end of the experiment by all groups.

-iv-Analysis of the Best Papers Written by Each Student

It is apparent that there was variation in the quality of the papers written by any given student over the course of the study. The mean S.D. of the ratings on the 112 sets of twelve papers was 6.59. The S.D. of this distribution of S.D.'s was 1.02. These values were obtained after pooling the standardized ratings over all five graders and finding a mean for each paper. The reliability of the summary Before-and-After scores is obviously affected by day-to-day variations in writing performance. It may be argued that writing performance could never exceed the limit of a student's hypothetical ability, but that many factors might cause a student to perform at less than his best. Since papers written on each of six consecutive class periods were pooled to form the summary scores, one or two aberrant scores due to "bad" days would make his mean performance unrepresentative of his ability. Following this argument, a new series of analyses was made with the summary Before-and-After rating scores.

First, each of the six scores in the Before-and-After blocks was summed over the five graders using the standardized ratings as the values. Then within each block the summed ratings on each paper were ranked in order of size. The poorest paper, *i.e.*, the paper with the smallest sum, was then eliminated from each of the blocks. The remaining five papers were then averaged for each block, and these values became the source data for a new analysis of variance. The entire process was then reiterated and summary values of the best four papers were used for the Before-and-After values. This process was repeated until only the best paper written before and the best paper written after the imposition of the independent variables were employed. The means and *F*-ratios for this progressive series of five comparisons are reported in Table 11. The analysis of variance results appear as Tables 21 through 25 in Appendix D. Inspection of Table 11 reveals no factors that change the interpretation of the results of the experiment from the analyses made earlier. With the exception of Section III (the group that graded papers of their peers during the middle ten weeks), the After means are less than the Before means for every comparison. Group 3 shows a very small improvement when the comparison is made for the best one, two, or three papers in the Before and After blocks of six papers. Only the Before-and-After *F*-ratio is significant at an acceptable level of confidence for all of the analyses. The increasing size of the means for each group as the poorer papers are eliminated from each comparison is noted. There appears to be no other trend in this table with the possible exception of a tendency for the size of the Methods *F*-ratio to increase slightly, as fewer, better papers are considered.

-v-*Analysis of the Diederich Factors of the Rating Sheet*

All analyses of the rating data to this point have used the total rating score on each paper. It will be remembered that the rating score was obtained by adding the values assigned to thirteen 5-point rating scales (see Appendix B). These thirteen items were grouped into five factors by the investigators that prepared the scale (10): *Ideas*, *Mechanics*, *Wording*, *Form*, *Flavor*. The sub-total scores from each rating sheet thus became the subject of the next

TABLE 11

Means and F-Ratios of the Best Papers Written Before and After the Experiment. Before-and-After Means Represent Summary of the N Papers Using Pooled Standardized Ratings.[1] N = 100.

	Best 5		Best 4		Best 3		Best 2		Best Paper	
Group	Before	After	Before	After	Before	After	Before	After	Before	After
1	10.04	9.98	10.26	10.20	10.36	10.31	10.46	10.42	10.60	10.50
2	10.13	10.10	10.38	10.31	10.50	10.39	10.65	10.50	10.82	10.64
3	10.06	10.00	10.27	10.24	10.35	10.37	10.47	10.49	10.61	10.69
4	9.92	9.79	10.15	9.98	10.26	10.08	10.38	10.16	10.53	10.29
5	10.01	9.83	10.19	10.02	10.29	10.10	10.40	10.21	10.53	10.32
F Section	.58		.69		.70		.79		.81	
F B&A	9.58**		9.53**		8.74**		10.75**		8.57**	
F Methods	.88		.83		1.25		1.68		1.85	

---

[1] For this analysis the z scores were converted to a standardized score with an M of 10.00 and a S.D. of 1.00.

**Significant at P < .01 for 1 and 95 d.f.

analysis of the study. Each of the five factors was separately studied to determine whether the conditions of the experiment might have contributed to improvement in any of these areas. It should be noted that the ratings for each factor were reduced to standardized values for each of the graders in a manner equivalent to that used with the total rating scores reported above. The means and standard deviations for each grader on each factor, employed for the z-score transformation, are given in Table 12. The mean and standard deviation of the raw ratings on each factor after over all five graders are also given in Table 12. This latter value was used to convert the z-scores to standardized equivalents of the raw data values. The mean Before and After values on each factor were obtained by summing the standardized factor values over all graders within the Before and After blocks of six papers.

The Ideas subtotals were obtained from the rating sheets, in raw score form, by adding the ratings given to four items: richness and soundness of ideas; development and support of ideas; relevance of ideas to the assignment; and adequate length without padding.

The factor labeled Mechanics was comprised of ratings of three items: grammar and usage; punctuation, capitals, abbreviations, numbers; and spelling.

The remaining three factors were represented by just two items each: Wording by choice and arrangement of words, and

TABLE 12

Means and Standard Deviations of Raw

Ratings for Each Grader for

Each Factor. N = 112.

	Ideas		Mechanics		Wording		Form		Flavor	
Grader	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.
1	11.64	2.99	8.57	2.07	6.18	1.32	5.78	1.57	5.34	1.77
2	10.82	2.10	7.11	2.30	5.85	.71	5.62	.94	5.91	.77
3	12.20	4.37	9.19	2.88	6.54	1.89	6.55	2.44	6.06	2.39
4	12.11	2.66	9.03	1.79	6.03	1.17	5.99	1.39	6.16	1.28
5	11.49	2.60	9.59	1.57	6.03	.86	5.38	1.26	5.25	1.38
Pooled	11.65	3.08	8.70	2.33	6.13	1.28	5.87	1.65	5.74	1.65

general level of vocabulary development; Form by organization (follows a discernible plan), and analysis (topic broken up into reasonable parts); Flavor by style (personal qualities of the writing), and interest and sincerity.

The means and *F*-ratios from the analysis of the data of each of the factors are given in Table 13. Analysis of variance tables for the five factors are given in Tables 26 through 30 of Appendix D.

The noteworthy elements of Table 13 are as follows. First, After means were lower than Before means with the single exception of the Mechanics and Wording factors of Section III. (The experimental activity characterizing Section III was practice in evaluating themes written by other students for ten weeks.) No Section or Methods *F*-ratio was large enough to permit rejection of the null hypothesis. All Before-and-After *F*-ratios were large enough to require rejection of the null hypothesis at the five per cent level of confidence; three of them require rejection at the one per cent level. The *F*-ratio for Methods in Wording approached significance at the five per cent level. Thus there is only the slightest suggestion that the analysis of factor means requires interpretation in any manner different from the analysis of the total ratings given earlier in Table 10.

-vi-Other Descriptive Analyses of the Data

It is apparent at this point that the hypotheses that directed the design of the study cannot be sustained by the

TABLE 13

Means and Fractions of Pooled Standardized
Summary Scores on the Five Sub-Sections
of the Rating Scale Total.

N = 100, Section N's = 20.

	Ideas		Mechanics		Wording		Form		Flavor	
Section	Before	After	Before	After	Before	After	Before	After	Before	After
1	11.58	11.51	9.18	8.95	6.09	5.99	5.91	5.90	5.67	5.59
2	11.84	11.77	9.10	8.96	6.29	6.17	6.06	5.98	5.92	5.86
3	12.34	12.07	8.76	8.81	6.14	6.25	6.00	5.94	5.97	6.04
4	11.43	10.73	8.23	8.20	6.02	5.97	5.73	5.50	5.58	5.30
5	11.65	11.00	8.53	8.24	6.17	6.01	5.82	5.69	5.78	5.56
Total Sample	11.77	11.42	8.76	8.64	6.14	6.08	5.90	5.80	5.78	5.67
F for Sections	1.27		1.56		.40		.84		1.50	
F for B&A	11.46**		3.28*		3.43*		4.70**		4.90**	
F for Method	1.78		.95		2.09		.64		1.38	

**Significant at less than the 1% level with 1 and 95 d.f.
*Significant at less than the 5% level with 1 and 95 d.f.

data of this experiment. Nevertheless, there remain several descriptive summaries of the results that are of interest: several correlations among the objective tests and the writing samples; the reliability of the five graders for both ranking ratings correlation of writing samples with objective test scores, and correlation of ratings with rankings; and finally the relationship of the scores on the writing samples to the topics assigned and to the temporal order in which they were written during the course of the experiment.

-vii-Correlations among the Objective Tests Scores and the Writing Samples

The four dependent variables of the experiment were CEEB scores, COOP scores, summary ratings of written compositions, and summary rankings of written compositions. These data were all obtained both before and after a ten-week experimental period (except for Section I, which had no intervening period). The issue of the relationship that exists among different methods of measuring composition ability is a live one--especially regarding the value of obtaining scores directly from writing samples as opposed to the more reliable and convenient objective tests. The correlations found in Table 14 throw some light on this question. This table also shows

TABLE 14

Intercorrelations among the Rating Sheet Factors and the Objective Tests. Correlation Coefficients on the Diagonal Represent Correlation of the Before with the After Values.

	After Values						
	Rating Sheet Factors					Obj. Tests	
	Mechanics	Wording	Form	Flavor	Total	Mean CEEB	Mean COOP
Ideas	.63	.79	.86	.92	.84	.53	.61
Mech.	.97	.80	.75	.73	.74	.55	.64
Word.		.94	.82	.92	.90	.64	.67
Form			.94	.87	.90	.54	.57
Flavor				.93	.94	.55	.65
Total					.97	.61	.69
CEEB						.61	.63
COOP							.67

the intercorrelations of the sub-sections of the rating scale with each of the other sub-sections and with the total rating score.[1]

It is interesting to note that the correlations of the total ratings with the two objective tests and the correlations of the tests with each other were all in the range of .61 to

---

[1] Each of the sub-section totals was removed from the total ratings before this correlation between sub-section and total was computed. The correlation of total ratings with other values in the table used all of the sub-sections.

.69. In other words the writing samples apparently reflected about as much, or as little, of the factors measured by each of the two objective tests as the two objective tests reflect of each other. The coefficient of determination of these three correlations ranges from .37 to .48.

Next it is useful to look at the correlations of the rating factor with each other and with the total ratings and the objective tests. We find that wording, form, and flavor are not as independent of total ratings as are ideas and mechanics. This may mean either that the graders were better able to discriminate ideas and mechanics and abstract them from any over-all impression they had of the paper or that in the constellation of factors that make for writing proficiency in students, these factors are more independent of the others. The generally high intercorrelations among the subtests suggest that the rating sheet in use does not separate the factors nearly as cleanly as may have been anticipated from the study by Diederich, et al. (10), where, in fact, they were the result of factor analysis techniques. Finally, none of the factor values correlate highly with the objective tests, but they all correlate moderately well with no dramatic differences. The range of correlations extends from the ideas vs CEEB r of .53 to the wording vs COOP r of .67.

The major diagonal of Table 14 represents correlations between measures of the same variables before and after the experiment. In view of the lack of significance, found in the analysis of variance, that might be attributed to treatment methods, these values may be considered to be reliability coefficients. It is interesting to observe that the reliability of the pooled graders' ratings, under the conditions of this experiment, were markedly higher than were the objective tests, _i.e_. r = .97 for total ratings vs. r = .61 and .67 for the CEEB and the COOP respectively. The sub-section reliability coefficients were remarkably large: .97 for ideas, .97 for mechanics, .94 for wording, .94 for form, and .93 for flavor. Thus the essentially negative results of the experiment can scarcely be attributed to lack of reliability of the graders' marks. It should be remembered that the time interval between the Before and After collection of writing sample data was considerably less than the interval between the administration of the two forms of the objective tests (see Procedure).

Thus the correlations of writing samples obtained before the experiment with the samples collected after are not entirely comparable to the correlation between the Before and After objective tests.

-viii-Grader Reliability

The temporal reliability of the graders in assigning scores to the writing samples was investigated for both the rating and the ranking methods. The graders rated the writing samples by going through the set of 112 papers one topic at a time. In each set of papers, six of the writing samples that appeared early in the set were selected and duplicates of these papers were inserted in the set toward the end. Thus each grader judged seventy-two papers, using the rating sheet, on two occasions, separated by the time required to grade the intervening approximately 90 to 100 papers. This time interval was not controlled, inasmuch as the graders worked independently throughout the summer of 1963. Their instructions were simply to score the papers in the order in which they received them and to finish a set on one topic before beginning another. They were not told which papers were duplicates. However, reports from the graders after they had completed their work indicated they recognized having read some of the duplicate papers earlier. There was no control by the experimenters over comparison by the grader of the rating sheet of the duplicate with the ratings given to the original copy, although they were requested to work on each paper individually. The set of intercorrelations for the seventy-two papers for the ratings for each judge is given in Table 15. The rather marked differences among the graders is found in the range of correlations for the total ratings of .56 to .96. In general the correlations were higher in the factors of ideas and mechanics than they were of the other three factors. These higher correlations may represent the greater variance possible in these two factors than in the others.

TABLE 15

Intercorrelations of Total Ratings and Sub-total Rating Factors of Seventy-Two Papers with Ratings Assigned upon Re-Grading by Each Grader

Factor	Grader				
	1	2	3	4	5
Ideas	.76	.95	.42	.90	.74
Mechanics	.73	.94	.68	.93	.72
Wording	.48	.93	.61	.89	.49
Form	.52	.87	.56	.90	.68
Flavor	.70	.92	.53	.81	.60
Total	.73	.93	.56	.96	.77

Reliability coefficients for each of the graders were determined for the ranking method also. It will be remembered that in ranking the writing samples, the graders worked with 112 sets of 12 papers each. They finished one set of papers on all topics for one student, before proceeding to the next set. Six duplicate sets of papers were given to each grader toward the end of the ranking task, representing six students whose papers were among the first to be ranked. There was, therefore, more time between the ranking of the duplicates and the ranking of the originals than was true of the ratings. This time was of the order of several weeks, whereas the time between ratings was of the order of a few days, although again there was no control over the speed of work habits of the graders. The six sets of duplicates were paired with the original papers and correlation coefficients obtained. These results are found in Table 16. The ranking

TABLE 16

Mean Rank-Order Correlation of Six Sets of Twelve Papers Ranked on Two Occasions by Each Grader. Fisher's z Transformation Was Used in the Computation.

Grader	Mean Correlation
1	.61
2	.97
3	.25
4	.83
5	.63

reliabilities are lower than the rating reliabilities. This reflects both an increased difficulty in making intra-writer discrimination as opposed to inter-writer judgments and the narrower ranges of the rank values as opposed to the ratings. Another check on the reliability of the graders is found in the correlation of the ranking given the twelve papers of each student writer and the ratings given the same papers. The correlations were converted to Fischer's z and the mean correlation for each determined. These data are presented in Table 17. It will be noted that the correlations for all graders are very close to .70 except for Grader Three whose mean r dropped to .47. The variabilities of the five sets are all very similar.

TABLE 17

Mean Correlations between Ratings and Rankings for the Same Papers as Judged by each Grader.

Grader	Correlations	S.D. of r (Fisher's z)
1	.68	.57
2	.71	.58
3	.47	.57
4	.70	.59
5	.69	.51

-ix-Mean Ratings per Order and per Topic

All papers by all students were sorted on the basis of the order in which they were written, *i.e.*, first, second, etc., and the mean raw ratings assigned by all five graders were computed. These results are presented in Table 18 and Figure 2. It is apparent in this presentation that there was an overall decline in writing ability from the first six papers to the last. The nature of the curve within the first six and last six papers cannot be said to be a reliable trend, however, due to the large variation within a given order compared to the differences between orders. The figure somewhat exaggerates the differences among orders

TABLE 18

Mean Ratings Per Order Assigned by All Graders to All Papers Written by All Students

Temporal Order	Mean	S.D.
1	38.31	6.05
2	39.40	6.30
3	38.46	6.17
4	37.90	6.33
5	38.32	6.10
6	38.24	6.45
7	38.51	5.78
8	37.63	6.72
9	37.43	6.38
10	37.68	6.20
11	37.95	5.82
12	37.21	6.21

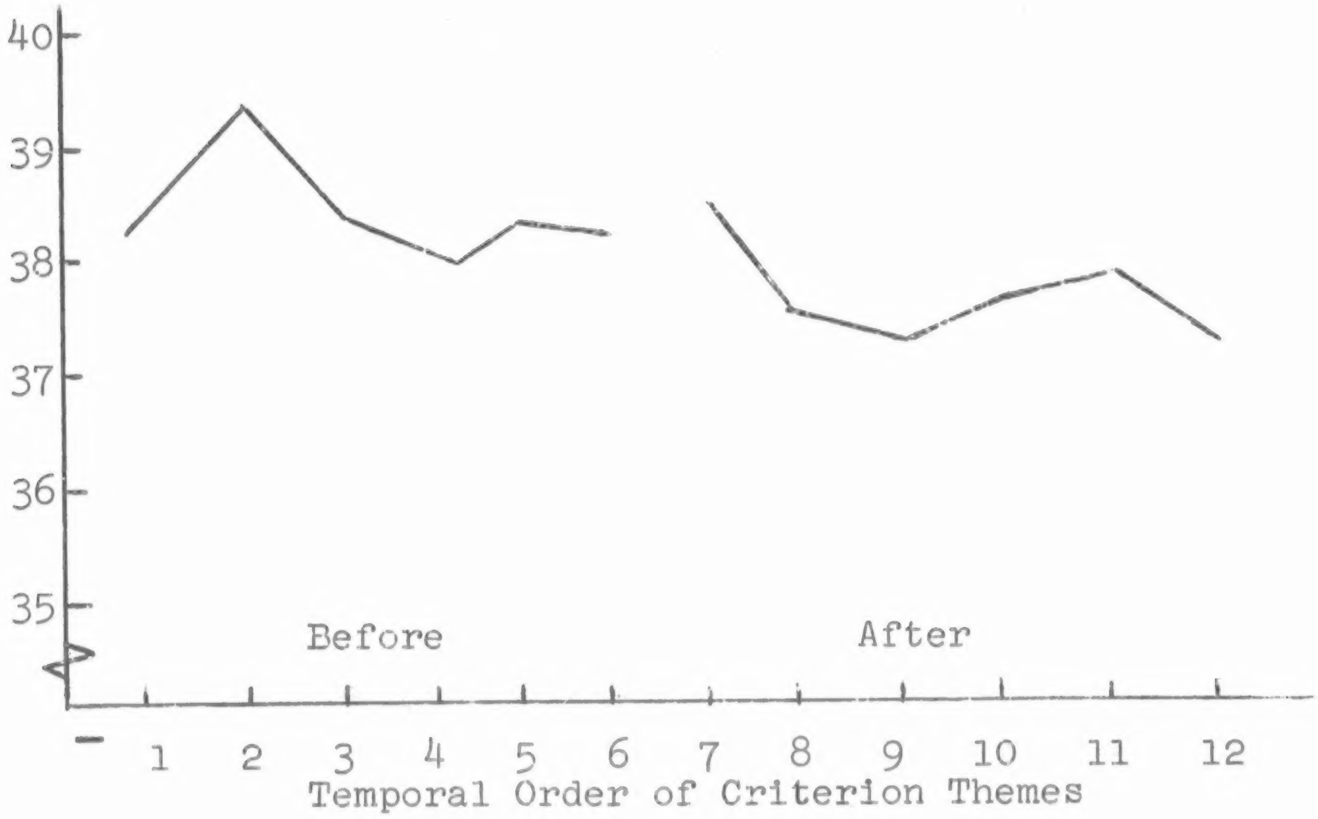

FIGURE TWO

Line Graph Showing General Decline in Theme Rating Over the Semester. Note the Shape of the Curve Within Each Block of Six Themes.

because of the necessity of foreshortening the ordinate in the graphic presentation.

The mean rating values among the various topics assigned were also computed and presented in Table 19. It will be remembered that the topics were presented in a different order for the students within any experimental section. Every topic appeared in all temporal orders of presentation over the complete set of students. The range of means and standard deviations is comparable to the values obtained by orders. Again the differences among topics were small relative to the variances within orders.

TABLE 19

Mean Rating by all Graders on all
Students for each Topic

Topic	Mean	S.D.
1	38.43	6.10
2	36.77	6.14
3	38.28	6.12
4	38.46	6.13
5	39.28	5.69
6	39.01	6.38
7	36.57	6.80
8	38.42	6.08
9	37.45	6.76
10	37.87	6.24
11	38.69	5.91
12	37.82	5.72

CHAPTER IV

DISCUSSION AND CONCLUSIONS

Support for the three empirical hypotheses advanced for this study must rest upon the F-ratios found in the interaction effect labeled Methods in the analysis of variance of the four dependent variables (*i.e.*, the two objective tests and the two measures from the writing samples). In no case was the Methods F large enough to allow rejection of the null hypothesis at a respectable level of confidence. The original hypotheses therefore stand without support as a result of this experiment. A number of *post mortem* comments seem in order, however. Some of these observations concern the capturing of the data in the Before and After periods, and some relate to the nature of the ten-week interval in which the experimental treatments were imposed.

-i-Advantages and Disadvantages of Obtaining Twelve Writing Samples

A distinctive feature of this experiment is the number of writing samples obtained from each subject, the number of graders employed, and the number of times each grader read each paper. Particularly important is the fact that writing performance was assessed before and after the impostion of experimental treatments by a block of six papers from each student. The stability that is possible in making inter-individual distinctions in writing performance under these conditions is indicated by a correlation of .97 between the mean ratings of the six papers written before and the mean ratings of the six papers written after the experiment. Each mean in this case represents the pooling of five raters on the six papers for each student. The score on each paper represents the total of thirteen rating scales. The correlation of Before with After means of the sub-total rating factors that included only two rating scales all were .93 or better (Table 15). Therefore the failure to find statistically significant results in the analysis of variance of these subjective scores can scarcely be attributed to unreliability inherent in the data. Unfortunately, it appears that the very procedure necessary to secure such stability introduced other factors that may have had a deleterious influence upon the results.

Campbell and Stanley have called attention to "external validity" (_i_._e_., generalizability of the results of an experiment) of experimental designs that employ a pretest.

> The effect of the pretest upon X [the independent variable] as it restricts external validity is of course a function of the extent to which such repeated measurements are characteristic of the universe to which one wants to generalize . . . . Where highly unusual test procedures are used, or where the testing procedure involves deception, perceptual or cognitive restructuring, surprise, stress, etc., designs having unpretested groups remain highly desirable if not essential. (12, p. 188)

It will be recalled that collecting the six Before and six After papers required all subjects in the experiment to write a paper during each of six consecutive class periods over two academic weeks at the beginning of the experiment and again for six periods after an intervening ten-week experimental interval. A most significant aspect of this procedure as seen by the student was that none of the papers were returned and no comment was made about them. In terms of usual English composition classes this does appear to be a highly unusual testing procedure. The repetitive, prolonged, and non-reinforcing character of this data-collecting period can scarcely have failed to influence the motivation and attitudes of the subjects. Unfortunately no systematic evaluation of these non-cognitive effects was obtained. Several comments from the instructors who proctored the collection of the writing samples suggest that expressions suggesting boredom and impatience were not uncommon during this period. The fact that the papers were written on assigned topics probably also served to reduce student ego-involvement with the task.

If, indeed, future research discloses the sensitivity of writing samples to the non-cognitive variables suggested here, the design of experiments on composition will need to move to a new level of sophistication in order to achieve reliability of writer performance, while at the same time avoiding the dampening effect of protracted testing.

-ii-The Value of Ranking vs. Ratings

The nature of the ranking data prohibited many comparisons with the other three forms of data collected. It will be remembered that the ranks were intra-individual comparisons over twelve papers whereas ratings and objective tests were essentially inter-individual comparisons. The conditions under which the judges read the papers do not permit an analysis of the relative ease of making intra-individual discriminations as opposed to inter-individual discriminations (i.e., for the ratings, the graders read twelve sets of 112 papers, each set consisting of a single topic for all students; for the rankings, the graders read 112 sets of twelve papers, each set consisting of all the papers written by a single student). The fact that the rating data failed to show the over-all decline in writing ability found in the ranking data suggests, however, that intra-individual distinctions are not as sensitively made as inter-individual distinctions. It is, of course, generally known that the size of intra-individual differences is not as large as inter-individual differences, even in a highly selected population such as a college freshman class.

-iii-The Rating Sheet Factors

No effort was made in this study to investigate the thirteen individual items on the rating scale. The analysis was confined to the five factors that comprise the total scale and the total rating values. Recent research has shown that these factors on this rating scale may be reduced to two major factors when applied to writing samples on ninth and tenth-grade students. (19) These have been identified as Content, associated with Ideas, Form, Flavor, and to some degree, Wording; and Technical Accuracy chiefly associated with Mechanics, and somewhat with Wording. When the five factors were independently subjected to analysis of variance in this experiment, no important differences in the results were discovered. There is the interesting observation, however, that the F-ratios for the Before-and-After effect in the Mechanics and Wording factors were smaller than they were for the other factors (Table 13). It is also interesting to note that the F-ratio for Methods comes closer to approaching significance in the Wording factor than for any of the other sub-sections of the rating scale. Finally, it is observed that the only means that did not show a decline from the Before to the After testing with the writing samples were the factors Wording and Mechanics for Section III. Future research may discover that these factors are less susceptible to the effects of motivational decline than other factors which are more influenced by the non-cognitive variables. It may also be true that the nature of the task give Section III (i.e., editing papers written by peers) contributes more to a Technical Accuracy factor than to Content. No attempt is intended by the above discussion to fly in the face of the results of the experiment in suggesting the presence of effects when the null hypotheses could not be rejected. There do appear to be external logical grounds for calling attention to these data.

-iv-The Objective Tests

The objective test results failed to disclose a decline in scores from the Before to the After testing; rather all sections showed what appeared to be an increase in mean scores. The assumption of non-cognitive influences on the writing samples is consistent with this reversal in the direction of the difference between the Before and After objective test means in terms of recent research by Rosner (19). He found that factor analysis of *Cooperative English Tests* and *SCAT* Verbal scores with ratings from a scale very similar to the one used in this study failed to identify common test and rating scale factors. This, he feels, suggests real differences in the abilities being measured.

> It is possible, therefore, that objective tests of English expression are concerned purely with cognitive properties, while subjective appraisals are influenced to some degree by the writer's affective (motivational), stylistic properties, or to some small measure by the writer's motor skills (graphic representation, rhythmical patterns). (19, p. 9)

Rosner concludes that the answer to the question of which type of test to use--objective or subjective--must be, for now, both types. Certainly the differences in the results obtained in the present study when objective and subjective tests were used lend support to his conclusion.

One fact remains regarding the data from the objective tests that needs discussion despite any difference in what objective tests measure as compared to written tests. The unexpectedly low correlation between the Before and After testing with two forms of the two tests needs examination. The test-retest correlation for the CEEB was .61, and for the COOP it was .67 (Table 15). The difference between the forms of the tests has already been mentioned. In the case of the CEEB it should be noted that the students took the first test at some time prior to, and for the purpose of, applying for admission to the University. For most students this may have been as much as a year prior to taking the After test at the end of their freshman year. For some students this time interval was even longer. The After test was not, of course, in any way related either to admission to college or to their academic standing. There was, therefore, a considerable difference in the perceived stress attendant upon taking the two tests. If the observations made earlier regarding boredom in collecting the writing samples are valid, it is likely that taking the second test was far from being an anxiety arousing situation. A marked interaction between general anxiety level, as a personality attribute, and situational stress is well known in experimental psychology. That is, subjects with high anxiety levels do relatively better under conditions of no stress than they do under threatening circumstances, whereas subjects with low anxiety levels do better under the motivating influence of "pressure." Since the subjects of this experiment were randomly chosen from the freshman class, it is reasonable to assume, even in the absence of supporting data, that there were in every section students at various points on the general anxiety level distribution. Perhaps student differences in conjunction with the different motivational properties between the Before and After testing situations plus the time interval between the sessions are sufficient to account for the relatively low test-retest correlation obtained with these well-established objective tests.

The COOP was administered at the beginning of the school year to all freshmen for the purpose of educational placement. There was not the same time interval nor the same degree of pressure attendant on the Before testing here as there may be assumed to have been with the CEEB. Nevertheless, the general argument appears to apply to this test as well as to the CEEB test. To be sure, the COOP test-retest coefficient is slightly higher than the CEEB correlation.

-v-Differences Among Graders

That there were important differences among the graders is apparent from the results of this study. The two graders who appear to have been most dissimilar were Graders Two and Three. Grader Two was a senior male faculty member and Grader Three a junior male faculty member. Graders One, Four, and Five were female, with Four being senior and One and Five graduate assistants. The S.D.'s of the total ratings varied from 4.98 for Grader Two to 12.44 for Crader Three (Table 8). There was also a consistent difference in the mean ratings assigned by those two judges, with Grader Three giving generally higher rates throughout the rating scale and Grader Two giving generally lower rates than the other judges. Marked differences were also found in rating-rerating reliability coefficients. Grader Two obtained an r of .93 for total ratings and Grader Three an r of .61. The difference was even more marked on the ranking-reranking reliabilities. Grader Two received a truly remarkable r of .97 whereas Grader Three obtained an r of .25.

These results suggest that selecting graders with proved reliability may be of more value to the efficiency of a study on written composition than the increase in the number of graders employed. In view of the pre-testing of the graders employed in this study, this also points up the need for long-term situational-type tests of grader consistency as opposed to the short "maximum performance" type of evaluations of graders. It seems unquestionably true that additudinal and motivation factors enter into grader performance as well as into writer performance.

-vi-The Experimental Treatments

The three experimental treatments assigned to Sections Three, Four, and Five were as follows: Section Three read and rated four peer writing samples each week and compared their rates with those assigned by three other members of the same section. Sections Four and Five wrote a theme a week, had the rated papers returned, and rewrote. The papers of Section Four were rated by peers; the papers of Section Five were read by a faculty member. The entire experimental period lasted ten weeks.

In retrospect, and in view of the evidence that none of these treatments can be said to have influenced writing performance as measured in this experiment in any significant way, two observations seem in order. First, the ten-week period was probably too short to change dramatically the writing habits of young adults who already write well enough to merit admission to the regular classes of a liberal arts college. It would seem to be a fairer test to impose the experimental treatments over the period of time equal to the entire freshman year. Second, in the effort to control the nature of the instruction involved in the treatment all instruction from the teacher in the classroom was forbidden. All information regarding composition practices had to be obtained from inspection of the corrected papers or from comparison of rating sheets assigned by different graders to the same paper. There were no didactic instruction, no textbook assignments, and no dialogue between students and professors in any of the sections that directly related to composition.

It appears now that a much more aggressive pedagogical posture should have been assumed in each of the experimental treatments. Such a procedure would have made it much more difficult to describe exactly the nature of the experimental treatment, and therefore more difficult to generalize the results had significance been found. Nevertheless, a more desirable research stratagem might be to establish experimental conditions that would maximally favor the possibility of a significant effect. Then, if the effect be found, future research may determine the functional relationships between characteristics of the treatment and the size of the effect.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. ANDERSON, C. C. The new STEP Essay Test as a measure of composition ability. *Educ. and Psych. Measurement*, 1960, 20, 95-102.

2. BRADDOCK, RICHARD, RICHARD LLOYD JONES, and LOWELL SCHOER. *Research in written composition*. Champaign, Ill.: National Council of Teachers of English, 1963.

3. BUXTON, E. W. An experiment to test the effects of writing frequency and guided practice upon students' skill in written expression. *Dissertation Abstracts*, 1958, 19, 709 (abstract).

4. CAST, B. M. D. The efficiency of different methods of marking English composition. *Brit. Journ. of Educ. Psych*, 1939, 9, 257-269; and 1940, 10, 49-60.

5. COLBERT, TAYLOR. Methodology in research in composition. *College Comp. and Comm.*, 1961, 12, 39-42.

6. COMMISSION ON THE ENGLISH CURRICULUM (NCTE). A check list for evaluating the English program in the junior and senior high school. *Eng. Journ.*, 1962, 51, 273-282.

7. CORBETT, EDWARD P.J. Do it yourself. *Coll. Eng.*, 1961, 22, 507-508.

8. DIEDERICH, PAUL B. The measurement of skill in writing. *School Rev.*, 1946, 54, 584-592.

9. -----. The Rutgers plan for cutting class size in two. *Eng. Journ.*, 1960, 49, 229-236.

10. ----- *et al*. *Factors in judgments of writing ability* (Research Bulletin 61-15). Princeton: Educational Testing Service, 1961. (Cited by permission.)

11. EDWARDS, ALLEN L. *Experimental design in psychological research*. New York: Rinehart, 1960.

12. GAGE, N. L., ed. *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

13. GRISSOM, LOREN V. Student leadership in evaluating composition. Eng. Journ., 1959, 48, 338-339.

14. HEYS, FRANK. The theme-a-week assumption: a report of an experiment. Eng. Journ., 1962, 51, 320-323.

15. HUDDLESTON, EDITH M. Measurement of writing ability at the college entrance level. J. Exp. Ed., 1954, 22, 165-213.

16. JONES, LYLE V. Some invariant findings under the method of successive intervals. In Psychological Scaling: Theory and Applications. New York: John Wiley, 1960, pp. 7-20.

17. MAIZE, RAY C. A study of two methods of teaching English composition to retarded college freshmen. Unpublished Ph. D. Dissertation, Purdue University, 1952.

18. MCGUIRE, EDNA. College freshmen on writing in high school. Eng. Journ., 1962, 51, 256.

19. ROSNER, BENJAMIN. Improving competence in English expression (Research Project A59-62, Progress Report No. 2). New York: Brooklyn College Office of Testing and Research, 1963.

20. SKINNER, B. F. Verbal behavior. New York: Appleton-Century-Crofts, 1957.

21. SMITH, DORA V. Class size in high school English: methods and results. Minneapolis: University of Minnesota Press, 1931.

22. STALNAKER, JOHN M. The construction and results of a twelve-hour test in English composition. School and Society, 1934, 29, 218-224.

23. STRICKLAND, RUTH. Evaluating children's composition. Element. Eng., 1960, 37, 321-330.

24. Theme reading and grading. Coll. Comp. and Comm., 1958, 9, 159-161.

25. WISEMAN, STEPHEN. The marking of English composition in grammar school selection. Brit. Journ. of Educ. Psych., 1949, 19, 200-209.

APPENDIX A

THEME TOPICS

The twelve theme topics are given below in the form in which they were presented to the students.

TOPIC 1

The Well-Rounded Versus the One-Sided Man

One of the goals of secondary-school and college education in America today is to produce well-rounded citizens--men and women with some knowledge of science and humanities, of the fine arts and civic affairs, of the past as well as the present, and of world as well as national or local concerns. Ideally, perhaps, all people should have such a balance. But turn your mind back a few years to a school or camp which you used to attend, a neighborhood you once lived in. Whom do you remember? Who made the vivid impressions on you? Was it the polished, well-balanced person, or was it the unusual, perhaps even eccentric man or woman with a peculiar but memorable one-sidedness? Is, in fact, "well-rounded" a synonym for "conforming" or "stereotyped"? Is balance of knowledge and personality gained only at the expense of individuality and dynamic power?

Discuss this problem in an essay, using as examples people both from your own experience and from literature you have read.

TOPIC 2

Steady Dating Is a New Social Evil

Students of social conditions among teen-agers notice with considerable alarm the tendency of high school and college students to limit their attentions to only one member of the opposite sex rather than taking a more "social" attitude at parties, dances, etc., and being part of a larger group of boys and girls. Some argue that this tendency is a result of a loss of traditional polite conduct and a trend toward the isolation of the individual. They feel, too, that part of the maturing process is learning to behave knowledgeably in mixed company and that the steady-dating principle defeats this end. Is there a defense for steady-dating among teen-agers? Write an essay in which you state that defense and either support it or attack it.

TOPIC 3

The Desirability of Grades in College

There are many arguments against the grading system in college--for example, subjectivity on the part of professors, distortion of real goals, improper emphasis on the merely intellectual side of education. On the other hand, it would seem that there must be some way to evaluate a student's performance, not only so that others will know how well he has done, but so he will know, himself. Write an essay in which you take one side or the other of this question and come to a reasoned conclusion of your own.

TOPIC 4

Does College Really Satisfy a Thirst for Knowledge?

A recent document that shocked the world of education stated that the college experience has absolutely no effect on students' values: that is, that they are not inclined to change their minds about important matters (political, social, religious, etc.) as a result of four years of college life. Other recent studies claim that not much really "happens" in college classrooms, that students forget all the facts, even most of the skills, like trigonometry and literary analysis. What, then, does a college education offer the student? Is it just a place where boys and girls can gracefully become men and women? For example, Cardinal Newman, author of the famous _The Idea of a University_, claimed that the purpose of education is to create "the gentleman." Or are there more tangible or non-tangible advantages to be gained? One might approach this problem by first classifying the kinds of students there are and then describing the goals they might achieve, or by working the other way around and classifying goals and then describing the kinds of students who would seek them.

In any case, write an essay in which you analyze the problem of the value of college education, coming to some reasonable conclusion of your own.

TOPIC 5

College Students Who Earn All or Part of Their Way Get More Out of College

We live in a culture that is incredibly rich. Furthermore, we have developed social and economic methods for taking care of all kinds of financial problems (medical, old-age, insurance against accident and death built into our salaries and taxes) which at one time the individual himself had to provide against. Finally, American parents are notoriously protective and indulgent. Where does the adolescent in our society learn self-reliance? Perhaps a better question would be, "Is it necessary for the adolescent in our society to learn self-reliance at all?" Write an essay in which you classify the life-situations in which such a virtue would be either desirable or necessary and describe how the student who "works his own way" will be better or worse equipped in that virtue.

TOPIC 6

Dormitory Rules for Men and Women Should Be the Same

There was a time when women were not even considered educable. Now, however, many educators (including assorted deans and counsellors) insist not only that women should be educated, but that they should take an increasingly important part in the political, economic, religious, and social life of the nation. Nevertheless, these educators are reluctant to give women the kind of complete freedom and responsibility they are willing to give college men. They argue that parents would react unfavorably, that the girls themselves would be unhappy because, as one girl put it, "When you've got a deadline, the boys _have_ to take you home whether they want to or not." Is there really a difference between boys and girls that requires treating the latter as a special class needing special protection? Write an essay in which you discuss this problem and come to a reasoned conclusion about it.

TOPIC 7

Do Television Programs Need Improvement?

The total number of hours per years spent by Americans watching television suggests the importance of television programs in our national culture. The range covered by the magic eye is enormous, including symphony concerts and the United Nations as well as sports events, westerns, and cartoons. And the commercials themselves vary enormously. Is such a wide range good? Are some types of programs unnecessary or even undesirable? Can standards be formulated by which the networks might voluntarily--or under legislative pressure--improve the quality or programs? Or is the quality already adequate?

Making use of specific illustrations drawn from well-known television programs, show what you believe needs to be done or why, in your opinion, no change need be made.

TOPIC 8

How Old Should Young Drivers Be?

The minimum age for a driver's license in most states, but not all, is sixteen. In view of the disproportionately large percentage of accidents caused by younger drivers, a number of states have considered raising the age requirement. Like most proposed reforms, this one has friends and enemies. The opinion of the great majority of teen-agers themselves is obvious. The opposing opinion of many older citizens is also obvious. The factors of transportation to school and work, parking and traffic problems, and social and moral patterns of behavior all enter into the question.

Giving clear and logical reasons for your position, with such illustrations or examples as may be appropriate, write an essay explaining your stand on this problem.

TOPIC 9

Facts May Change but Wisdom Remains

In a broad sense, the purpose of a college education is to make men wise. But this rather general level of wisdom is developed in college by a set of particular and specialized courses. In many of these courses, the student is expected to learn names, dates, facts, formulas, and so on.

Facts change, however. Many of today's college chemistry teachers, for example, when they were undergraduates, learned as facts things which are now known not to be true.

Historians have changed, during the past generation, some of their ideas and interpretations of modern history. Some books and authors important half a century ago are almost forgotten, and other worshiped in their places. Yet such changes do not mean that education is worthless, or that no truths remain constant.

Discuss the question of the relationship between facts and wisdom as it applies to your own education today.

TOPIC 10

Life in the City Versus Life in the Country

Proponents of city living have much to say in defense of their way of life: nearness to work; accessibility of doctors and hospitals; availability of public transportation. In addition they mention the cultural advantages of museums, libraries, symphony concerts. Despite the strength of these arguments, there are many people who prefer life in the country; and they speak of values perhaps less tangible in some respects but nonetheless real to them. The beauties of nature, good, clean fresh air, room to live in and the privacy that goes with it--these are some of the reasons they give. And of course there are other things to be said on both sides.

Choose one side or the other, and write a convincing essay showing why the way of life you have chosen is clearly the better.

TOPIC 11

The Value of Varsity Athletics in College

During the past half century, probably no phase of American university and college life has come under public criticism so constantly, or been attacked and defended with such heat, as athletics. With an intense athletic program come many obvious problems: large expenditures for equipment and facilities, coaches' salaries, and athletic scholarships; a special caste of students not primarily interested in--and often not qualified for--academic studies; the intrusion into college affairs of all sorts of outsiders including the press, sports-minded politicians, professional athletic organizations, and even professional gamblers and gangsters. There are other less obvious but equally significant effects on the nature and tone of campus life. On the other hand, university administrators and trustees are well aware of what varsity football can do in such matters as reputation, alumni support, and legislative appropriations.

Limiting your discussion to a single type of institution (*e.g.*, the large state university or the private liberal arts college) present a reasoned opinion of the value and desirability of varsity athletics.

TOPIC 12

Extra-Curricular Activities in College

It is a platitude that in college, studies come first; it is also a platitude that much college education takes place outside courses. Few students escape extra-curricular activities altogether. Some, on the other hand, devote most of their time to them. Which extra-curricular activities should be encouraged on a college campus? Which ones banned? Why? Is there any set of criteria for desirable activities by which a student could judge which and how many he ought to undertake, or by which an administrator could determine which were important and valuable?

Write an essay giving your views on this subject and supporting them by specific examples of extra-curricular college activities.

APPENDIX B

RATING SHEET USED IN THE EXPERIMENT

RATING SHEET Paper # ___ Reader # ___ Date _____

1 low 5%	2 next fifth	3 middle half	4 next fifth	5 top 5%
Failing	Passing	Good	Superior	Exceptional

IDEAS						
Richness and soundness of ideas	(1)	1	2	3	4	5
Development and support of ideas . . .	(2)	1	2	3	4	5
Relevance of ideas to the assignment .	(3)	1	2	3	4	5
Adequate length without padding	(4)	1	2	3	4	5
MECHANICS						
Grammar and usage	(5)	1	2	3	4	5
Punctuation, capitals, abbreviations, numbers	(6)	1	2	3	4	5
Spelling	(7)	1	2	3	4	5
WORDING						
Choice and arrangement of words	(8)	1	2	3	4	5
General level of vocabulary development	(9)	1	2	3	4	5
FORM						
Organization (follows a discernible plan)	(10)	1	2	3	4	5
Analysis (topic broken up into reasonable parts)	(11)	1	2	3	4	5
FLAVOR						
Style (personal qualities of the writing)	(12)	1	2	3	4	5
Interest and sincerity	(13)	1	2	3	4	5

COMMENTS: SUM ________

APPENDIX C

SUMMARY OF DATA FROM ALL STUDENTS

APPENDIX C

SUMMARY OF DATA FROM ALL STUDENTS

		BEFORE (FIRST SIX PAPERS)					AFTER (LAST SIX PAPERS)				
STD NBR	SEC NBR	CEEB	COOP	RW RATES	STAND RATES	RNKS	CEEB	COOP	RAW RATES	STAND RATES	RNKS
1	4	558	161	33.77	34.51	8.10	513	169	30.10	30.40	4.90
2	1	473	160	34.73	34.99	7.57	555	167	31.37	31.72	5.43
3	1	631	175	42.93	42.67	6.00	577	175	43.33	42.99	7.00
4	2	497	168	41.80	41.78	7.23	433	167	42.10	42.19	5.77
5	5	437	162	40.43	40.35	8.10	477	169	33.63	33.89	4.90
6	5	625	168	44.43	44.64	7.73	662	174	40.73	41.06	5.27
7	2	570	168	40.07	39.79	5.07	413	158	42.83	42.89	7.93
8	5	598	167	40.50	40.60	6.17	435	166	43.40	42.70	6.83
9	1		163	31.30	31.56	5.00	377	165	32.53	33.12	8.00
10	3	558	175	42.00	41.43	5.90	609	178	46.57	46.18	7.10
11	3	474	166	37.77	37.90	7.27	328	166	34.93	34.79	5.73
12	4	528	163	35.33	35.98	7.43	588	171	32.67	32.85	5.57
13	1	520	162	36.43	37.08	6.63	470	171	38.07	38.75	6.37
14	5	561	168	36.13	36.41	7.43	577	166	35.67	35.40	5.57
15	1	556	165	40.13	40.06	5.97	534	174	40.20	40.30	7.03
16	4	505	103	33.37	33.28	7.03	570	164	33.37	33.18	5.97
17	1	658	173	38.63	39.00	6.67	421	175	39.73	40.16	6.33
18	2	546	173	38.53	39.40	7.43	492	175	39.47	39.89	5.57
19	2		153	30.53	31.37	7.43	318	153	28.87	29.58	5.57
20	5	603	168	34.93	35.65	5.47	635	174	36.87	36.79	7.53
21	1	518	173	40.47	41.07	6.10	538	169	41.13	41.03	6.90
22	1	431	155	34.03	34.21	5.33	449	161	35.70	35.89	7.67
23	2	437	164	36.33	36.38	7.00	435	160	36.60	36.15	6.00
24	4	479	158	37.23	37.17	5.83	406	160	37.07	36.88	7.17

APPENDIX C CONTINUED

		BEFORE (FIRST SIX PAPERS)					AFTER (LAST SIX PAPERS)				
STD NBR	SEC NBR	CEEB	COOP	RAW RATES	STAND RATES	RNKS	CEEB	COOP	RAW RATES	STAND RATES	RNKS
25	4	550	159	35.47	35.60	7.97	521	168	28.93	29.39	5.03
26	4		160	38.10	39.12	5.27	442	162	38.53	39.80	7.73
27	1	347	162	31.97	32.36	6.13	350	158	30.53	31.21	6.87
28	1	526	161	36.13	36.58	6.80	451	168	33.37	34.06	6.20
29	3	728	185	48.77	48.16	5.90	626	188	52.50	52.10	7.10
30	3	518	171	38.00	37.75	6.17	435	171	38.20	38.08	6.83
31	5	598	172	46.43	46.14	7.10	563	179	43.67	43.18	5.90
32	4	467	160	33.73	34.24	5.87	413	162	34.90	34.95	7.13
33	1	437	166	40.57	40.80	7.07	503	165	40.57	40.57	5.93
34	5	532	168	40.20	39.74	7.33	520	166	39.33	39.04	5.67
35	2	655	180	44.63	44.97	6.10	555	183	43.97	43.48	6.90
36	3	534	169	41.83	41.52	6.33	499	176	41.00	40.73	6.67
37	2	443	167	39.83	39.95	6.63	529	171	39.13	39.10	6.37
38	3	389	163	43.13	43.02	7.13	534	166	40.17	40.85	5.87
39	4	532	184	50.50	50.50	7.17	605	183	48.40	48.43	5.83
40	2		174	38.03	37.43	5.13	527	174	41.63	41.95	7.87
41	1	498	167	44.30	43.79	5.47	520	164	45.33	45.13	7.53
42	1	728	186	49.50	49.43	7.63	719	184	46.67	46.99	5.37
43	3	516	166	41.80	41.79	6.97	456	171	41.87	41.24	6.03
44	4	604	174	43.97	43.91	7.20	506	178	42.33	41.63	5.80
45	5	544	160	37.30	37.48	6.93	477	175	37.80	37.72	6.07
46	4	395	163	32.13	32.59	7.73	497	164	28.53	29.32	5.27
47	5	534	171	41.37	41.23	8.27	445	163	34.33	34.79	4.73
48	2	389	178	40.03	40.15	6.60	371	169	39.20	39.03	6.40

APPENDIX C CONTINUED

STD NBR	SEC NBR	BEFORE (FIRST SIX PAPERS) CEEB	COOP	RAW RATES	STAND RATES	RNKS	AFTER (LAST SIX PAPERS) CEEB	COOP	RAW RATES	STAND RATES	RNKS
49	2	492	171	36.43	36.23	5.60	584	171	36.53	37.13	7.40
50	1	528	168	41.33	41.73	6.77	506	170	42.17	41.67	6.23
51	3	523	172	38.03	38.29	6.30	477	171	37.13	36.87	6.70
52	5	558	171	38.53	38.97	7.63	527	176	36.57	36.04	5.37
53	3	752	185	47.57	47.81	6.63	634	182	44.97	44.72	6.37
54	4	515	163	35.37	35.73	6.80	634	165	35.73	35.64	6.20
55	2	419	157	33.00	33.22	6.20	421	156	30.60	30.82	6.80
56	4	475	168	32.80	33.42	5.53	520	164	31.90	32.55	6.47
57	4	540	175	43.77	42.88	6.37	584	174	43.67	43.35	6.63
58	3	636	178	40.47	40.46	6.40	559	174	40.30	40.53	6.60
59	3	655	171	41.17	41.17	6.90	594	168	38.70	39.21	6.10
60	2	554	174	43.23	42.99	7.67	497	172	39.67	40.01	5.33
61	4	455	161	29.07	29.77	6.13	534	163	27.93	28.23	6.87
62	3	534	169	35.90	36.09	6.67	442	168	34.77	35.09	6.33
63	2	407	162	40.70	40.35	7.23	449	161	37.27	37.45	5.77
64	2	437	168	38.67	38.88	5.97	471	161	39.70	39.39	7.03
65	5	475	163	37.00	36.93	6.60	438	160	37.07	37.25	6.40
66	3	534	167	43.10	42.57	5.77	506	176	44.77	44.46	7.23
67	1	468	160	39.97	39.48	5.90	484	164	38.77	39.02	7.10
68	1	480	167	37.20	37.24	6.80	449	161	36.90	36.85	6.20
69	5	383	161	26.33	26.20	7.23	660	160	25.10	25.38	5.77
70	3	661	173	46.87	46.69	6.90	534	177	43.83	43.24	6.10
71	2	607	171	45.60	45.18	6.90	570	172	45.33	44.61	6.10
72	4	592	171	38.47	38.40	6.17	527	176	39.10	38.62	6.83

APPENDIX C CONTINUED

		BEFORE (FIRST SIX PAPERS)					AFTER (LAST SIX PAPERS)				
STD NBR	SEC NBR	CEEB	COOP	RAW RATES	STAND RATES	RNKS	CEEB	COOP	RAW RATES	STAND RATES	RNKS
73	3	652	182	45.90	45.89	7.73	570	181	39.37	39.31	5.27
74	1	595	172	41.07	40.53	7.13	634	178	42.23	41.37	5.87
75	5	623	171	41.40	41.55	7.07	614	174	40.30	39.77	5.93
76	1	431	163	29.53	30.27	6.67	406	166	28.33	29.11	6.33
77	3	455	168	38.07	38.09	7.60	470	168	37.37	36.79	5.40
78	5	494	170	34.87	34.88	6.93	426	160	32.77	32.93	6.07
79	3		139	22.00	22.18	6.17		137	23.73	24.03	6.83
80	4	617	170	33.67	33.93	6.77	485	175	34.07	34.06	6.23
81	3	431	168	37.03	37.74	5.67	392	168	38.17	38.24	7.33
82	4		164	38.13	38.33	6.90	484	165	37.10	37.22	6.10
83	2	580	173	43.27	43.37	6.53	541	178	41.90	41.94	6.47
83	5	401	151	26.10	26.65	6.43	359	156	28.63	29.16	6.57
85	2			41.80	41.46	6.33	515	173	43.50	43.85	6.67
86	3	573	171	46.77	46.38	7.80	605	176	41.93	41.82	5.20
87	5	468	147	34.10	34.56	6.50	484	163	31.50	31.82	6.50
88	1	538	171	41.27	41.07	7.77	591	170	37.90	38.03	5.23
89	1	492	162	35.90	35.67	5.87	520	163	35.93	35.90	7.13
90	4	610	179	44.40	44.53	4.93	622	175	46.67	47.07	8.07
91	5	640	169	31.97	32.59	6.53	616	172	33.23	33.52	6.47
92	1	652	176	45.63	45.04	7.17	601	184	43.53	43.29	5.83
93	3	631	172	44.33	44.45	5.90	492	174	43.43	44.05	7.10
94	2	514	163	39.80	40.60	7.53	598	158	37.00	37.50	5.47
95	1	371	153	35.43	35.45	6.33	364	163	35.33	35.47	6.67
96	3	468	170	35.33	35.83	6.83	428	169	35.47	36.31	6.17

APPENDIX C CONTINUED

		BEFORE (FIRST SIX PAPERS)					AFTER (LAST SIX PAPERS)				
STD NBR	SEC NBR	CEEB	COOP	RAW RATES	STAND RATES	RNKS	CEEB	COOP	RAW RATES	STAND RATES	RNKS
97	3	676	167	40.53	40.70	6.07	520	166	38.63	39.09	6.93
98	5	543	170	40.07	40.03	5.87	463	171	40.17	40.40	7.13
99	5	498	167	37.27	37.26	6.67	506	169	35.87	34.99	6.33
100	2	540	157	36.73	37.06	6.20	563	166	35.53	36.25	6.80
101	3	377	148	35.30	35.51	4.87	477	164	38.37	38.75	8.13
102	5	550	172	39.27	39.94	7.20	541	168	35.70	36.21	5.80
103	4	546	168	35.67	36.13	7.33	499	175	35.37	35.68	5.67
104	2	377	159	33.80	34.22	5.73	477	166	33.17	33.44	7.27
105	2	568	168	38.53	38.79	6.70	513	173	36.13	36.19	6.30
106	5	406	135	30.90	31.36	6.40	342	155	31.00	31.54	6.60
107	3		168	44.30	44.50	6.13	563	167	46.10	46.10	6.87
108	5	546	169	37.00	37.01	8.37	523	164	30.73	31.16	4.63
109	1	320	151	29.27	30.33	5.80	316	162	31.07	31.82	7.20
110	5	635	166	37.87	38.52	5.93	529	167	39.67	40.27	7.07
111	3	377	158	39.33	40.02	5.67	452	165	42.27	42.41	7.33
112	4	409	165	33.30	33.82	7.23	447	161	31.83	32.31	5.77

APPENDIX D

ANALYSIS OF VARIANCE TABLES NOT INCLUDED IN THE TEXT

TABLE 20

Analysis of Variance for the Mean Pooled Standardized Ratings on the Best Paper Written Before and the Best Paper Written After the Experiment

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	3.0	4	.7	.81
Error Between	87.3	95	.9	
Within Ss				
Before-and-After	.8	1	.8	8.59**
Methods (Sec X B&A)	.7	4	.2	1.85
Error Within	8.8	95	.1	
Total	100.6	199		

**Significance at P < .01 for 1 and 95 d.f.

TABLE 21

Analysis of Variance for the Mean Pooled Standardized Ratings on the Best Two Papers Written Before and the Best Two Papers Written After the Experiment

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	2.6	4	.6	.79
Error Between	78.1	95	.8	
Within Ss				
Before-and-After	.7	1	.7	10.75 **
Methods (Sec X B&A)	.4	4	.1	1.68
Error Within	6.2	95	.1	
Total	88.0	199		

**Significance at $P < .01$ for 1 and 95 d.f.

TABLE 22

Analysis of Variance of the Mean Pooled Standardized Ratings of the Best Three Papers Written Before the Experiment and the Best Three Written After the Experiment

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	2.2	4	.5	.70
Error Between	73.1	95	.8	
Within Ss				
Before-and-After	.5	1	.5	8.74**
Methods (Sec X B&A)	.3	4	.1	1.25
Error Within	5.6	95	.1	
Total	81.7	199		

**Significance at $P < .01$ for 1 and 95 d.f.

TABLE 23

Analysis of Variance of the Mean Pooled Standardized Ratings of the Best Four Papers Written Before the Experiment and the Best Four Written After the Experiment

Source of Variation	Sum of Squares	d.f.	Mean Square	_F_
Between Ss				
Sections	2.1	4	.5	.69
Error Between	71.0	95	.8	
Within Ss				
Before-and-After	.5	1	.5	9.53**
Methods (Sec X B&A)	.2	4	0.0	.83
Error Within	5.1	95	.1	
Total	78.8	199		

**Significance at $P < .01$ for 1 and 95 d.f.

TABLE 24

Analysis of Variance of the Mean Pooled Standardized Ratings of the Best Five Papers Written Before and the Best Five Papers Written After the Experiment

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	1.7	4	.4	.59
Error Between	67.2	95	.7	
Within Ss				
Before-and-After	.5	1	.5	10.04**
Methods (Sec X B&A)	.1	4	0.0	.68
Error Within	4.7	95	0.0	
Total	74.2	199		

**Significance at $P < .01$ for 1 and 95 d.f.

TABLE 25

Analysis of Variance of the Mean Pooled Standardized Ratings of the *Ideas* Factor

Source of Variation	Sum of Squares	d.f.	Mean Square	*F*
Between Ss				
Sections	30.3	4	7.6	1.27
Error Between	563.8	95	5.9	
Within Ss				
Before-and-After	6.2	1	6.2	11.47**
Methods (Sec X B&A)	3.8	4	1.0	1.78
Error Within	51.1	95	.5	
Total	655.2	199		

**Significance at $P < .01$ for 1 and 95 d.f.

TABLE 26

Analysis of Variance of the Mean Pooled Standardized Ratings of the Mechanics Factor

Source of Variation	Sum of Squares	d.f.	Mean Square	*F*
Between Ss				
Sections	23.4	4	5.9	1.56
Error Between	355.9	95	3.7	
Within Ss				
Before-and-After	.8	1	.8	3.29
Methods (Sec X B&A)	.9	4	.2	.95
Error Within	22.1	95	.2	
Total	403.1	199		

TABLE 27

Analysis of Variance of the Mean Pooled Standardized Ratings of the Wording Factor

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	1.6	4	.4	.40
Error Between	94.6	95	1.0	
Within Ss				
Before-and-After	.2	1	.2	3.43
Methods (Sec X B&A)	.5	4	.1	2.09
Error Within	5.2	95	.1	
Total	102.1	199		

TABLE 28

Analysis of Variance of the Mean Pooled Standardized Ratings of the Form Factor

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Between Ss				
Sections	4.3	4	1.1	.84
Error Between	121.9	95	1.3	
Within Ss				
Before-and-After	.5	1	.5	4.70*
Methods (Sec X B&A)	.3	4	.1	.64
Error Within	10.9	95	.1	
Total	137.9	199		

*Significance at P < .05 for 1 and 95 d.f.

TABLE 29

Analysis of Variance of the Mean Pooled Standardized Ratings of the Flavor Factor

Source of Variation	Sum of Squares	d.f.	Mean Sqaure	F
Between Ss				
Sections	8.0	4	2.0	1.50
Error Between	126.8	95	1.3	
Within Ss				
Before-and-After	.6	1	.6	4.90**
Methods (Sec X B&A)	.7	4	.2	1.38
Error Within	12.5	95	.1	
Total	148.7	199		

**Significance at P $<$ .01 for 1 and 95 d.f.